14–17 Nucleic Acids Research, 1997, Vol. 25, No. 1

DNA Data Bank of Japan in the age of information biology

Yoshio Tateno* and Takashi Gojobori

Center for Information Biology, National Institute of Genetics, Yata, Mishima 411, Japan

Received September 16, 1996; Revised and Accepted October 8, 1996

ABSTRACT

DNA Data Bank of Japan (DDBJ) began its activities in 1986 in collaboration with EMBL in Europe and GenBank in the United States. DDBJ developed a data submission tool called Sakura, by which researchers can submit their newly sequenced data on WWW from every corner of the world. The data bank also built a database management system (Yamato II), incorporating the techniques and functions of the object-oriented database, in order to efficiently process the data it has collected. A number of research activities in information biology are also going on at DDBJ. Two such activities are also briefly introduced in this report.

INTRODUCTION

The DNA Data Bank of Japan (DDBJ) began its activities in earnest in 1986 in collaboration with EMBL and GenBank. Before beginning, there was a series of discussions among Japanese molecular biologists and biophysicists about the organization in which the data bank should be established. The discussion finally resulted in a proposal for the data bank to be founded in the National Institute of Genetics (NIG), which is governed by the Ministry of Education, Sports, Science and Culture (MESSC). The proposal was soon implemented with the endorsement of MESSC. Since then, continuous support by MESSC has made it possible to maintain DDBJ activities, and establish a new center in 1995 at NIG. It is called the Center for Information Biology (CIB), which is composed of four research laboratories. The four laboratories devote themselves not only to operating DDBJ but also to their own research projects in information biology with a focus on molecular evolution.

The establishment of CIB perhaps reflects our awareness of the importance of a newly developed discipline, information biology. Incidentally, we prefer information biology to bioinformatics, because the former implies the placing of a heavier weight on biology than on informatics, while the latter does it the other way around. We strongly believe that the international DNA sequence databases (DDBJ/EMBL/GenBank) have greatly contributed to the development of information biology by providing researchers worldwide with DNA sequence data, and related information and software. It should be noted that the three data banks exchange the data they have collected and processed on a daily basis so that the three databases are essentially in synchronization with each other. The activities of DDBJ now include data collection, processing and dissemination and software development, as will be reported in the following sections.

DATA COLLECTION

There are two ways to submit sequence data to DDBJ, through Authorin and Sakura (1). The former was developed in GenBank to be provided on floppy disk, and the latter was recently devised in DDBJ to be operated on WWW. We expect that Authorin will soon be replaced by Sequin, which was newly developed in GenBank. As use of WWW has been spreading rapidly around the globe, however, we are particularly encouraging possible data submitters to utilize Sakura.

There are two versions of Sakura (http: //sakura.ddbj.nig.ac.jp) available at present, the Japanese and English versions. Both have the following six resources: (i) a page resource regulating the flow of the pages for data input in the browser; (ii) a form resource describing the type and number of data items on each page; (iii) a menu resource containing lists which will appear in the pop-up-menu and list-box; (iv) an error check resource for parsing the data typed in; (v) an item-dependency resource defining the inter dependency among the data items; and (vi) a word resource containing the names of the data items in different languages. These resources enable Sakura to be equipped with the gateway script with three major functions: (i) checking the data given by a submitter, referring to the error check and the item-dependency resources; (ii) creating pages in HTML, referring to the page, form, menu and word resources; and (iii) making a transaction file to install the data into the database.

With the first function in particular, submitters can make a rigorous error check on the data they have typed in before submission. Sakura issues the submitters the mandatory error when they have not given such items as the submitter's name and address, the date for releasing the data to the public, the source organism of the sequence, and the sequence data. If the data to be submitted describes a coding region, the information about the structure of the region and its product should also be given by following the way Sakura guides. If such information is not included, the sequence will be regarded as EST data (2) and treated as EST in the data processing and dissemination. This is now the common practice among the three international databases (DDBJ/EMBL/GenBank) in order to maintain the quality

* To whom correspondence should be addressed. Tel: +81 559 81 6857; Fax: +81 559 81 6858; Email: ytateno@genes.nig.ac.jp

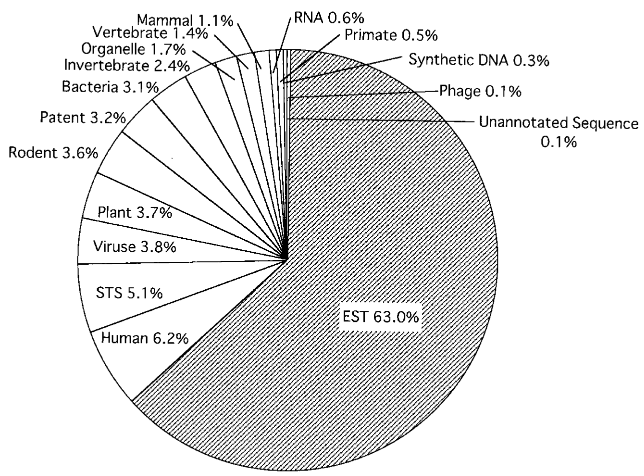

Figure 1. Proportion of categorized data submitted to the international DNA sequence databases (DDBJ/EMBL/GenBank). The proportion is represented in the percentage of the number of data submissions. The data were taken from the newest release by DDBJ (DDBJ release 26, July 1996). The total number of submissions for the release is 835 552.

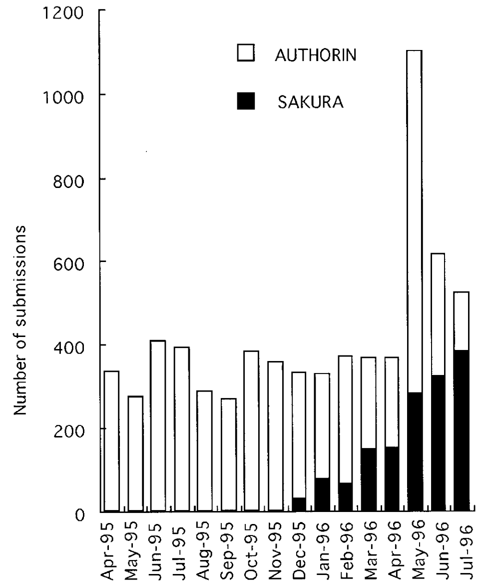

Figure 2. Recent data submissions to DDBJ. The white bar represents the monthly number of submissions through Authorin, and black bar shows the number through Sakura, which was made public in December 1995.

of the data they maintain and release to the public. Of course, most EST data were specified as EST beforehand by the submitters. The databases differentiate the ordinary sequence data from ESTs so that the users of the databases will not mix them up. As EST data submitted from ongoing genome projects accumulate at an enormous rate in the international databases, and now amount to 63% of the total data as shown in Figure 1, this practice will be increasingly crucial to the users.

In Figure 2 we show the recent statistics for data submissions to DDBJ. Since Sakura became open to the public in December 1995, the number of submissions through the tool has steeply increased. The most recent statistics in the figure indicate that 73% of the submissions were made through Sakura. Thus we are confident that most of the submissions to DDBJ will be made this way in the future, which encourages us to improve Sakura continuously. One way to improve it that we are now considering is that Sakura be displayed on the screen in other Asian languages such as Korean and Chinese. This can be made possible in collaboration with people in the corresponding countries by use of the word resource mentioned above. (The data themselves should of course be given in English only, no matter what languages the submitters choose.)

DATA PROCESSING

For nearly half a decade, we had relied on AWB (the Annotator's Work Bench) developed and kindly provided by the Los Alamos National Laboratory to process submitted data. Though the tool was quite useful, we were not able to make the modifications to it that our processing demands required. Thus we decided to develop a data processing system ourselves, and replaced AWB with the newly developed system in our data processing practice in January 1996. We named it Yamato II (3). Yamato II was written in C++, because it had to deal with many objects (tables) in the database and to be modified and extended further on demand. This computer language has advantages in incorporating these situations over other languages like C and Pascal.

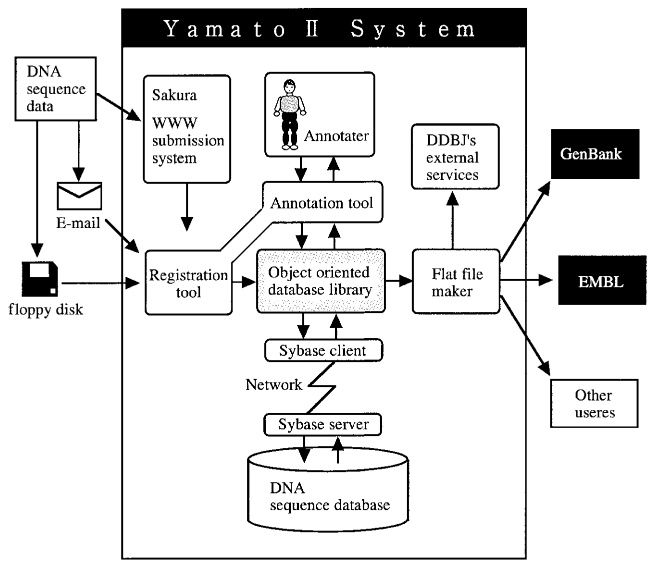

Figure 3. New data management system at DDBJ, Yamato II. This shows the data flow from data submission to data distribution in Yamato II. The core of this system is the object oriented database library in the shaded box.

The outline of Yamato II is shown in Figure 3. The registration and annotation (R & A) tools in this figure roughly correspond to AWB, implying that Yamato II covers greater database operations than AWB. The R & A tools receive the submitted data through Authorin and Sakura, and make primary checks on them, and issue an accession number to the submitter, if they have passed through the process without causing an error. These steps are not fully automatic, but also need experienced human interpretation, parsing and decision making. As to the accession number, the

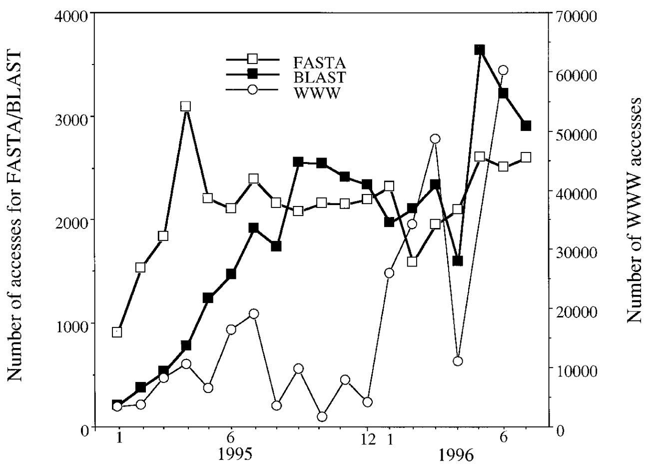

Figure 4. Number of accesses to DDBJ on WWW. The open square shows the monthly number of accesses to FASTA, closed square shows that to BLAST, and open circle represents the number of the other WWW accesses. For the numbers for FASTA and BLAST, refer to the left ordinate, and for the other WWW accesses, refer to the right one.

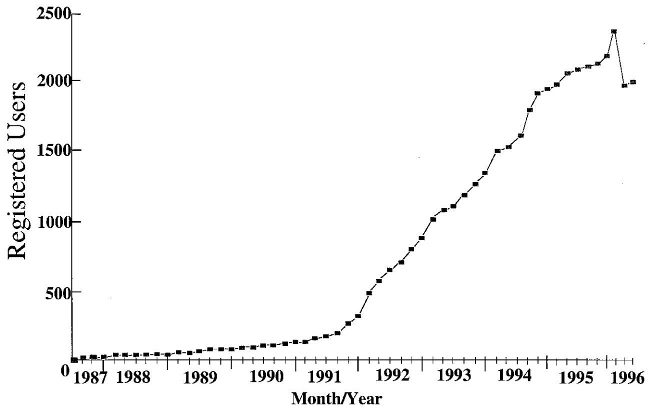

Figure 5. Number of users for DDBJ. The closed square shows the monthly number of the users for DDBJ.

three databases recently agreed to extend the present form to a new one, which is made up of two letters and six digits such as AB123456. This was of course due to the fact that the number of submissions had increased at an enormous speed, which gave the three databases the clear warning that they would soon use up all the possible accession numbers. This rate of increase is attributable mainly to the submissions of EST data from a number of ongoing genome projects, as mentioned above. Thus if one is to use or develop software which directly involves the accession number, one has to be aware of the new accession number and make proper adjustments or modifications.

A new aspect of Yamato II is that it includes our original object oriented database library, which functions as the interface between the R & A tools and the database built on Sybase. The object oriented library gives Yamato II the following three advantages: (i) development of applications in the R & A tools can be done independently of the structure and assimilation procedures of the database; (ii) since the descriptions of the objects in the library make a one-to-one correspondence to the definitions in the database, the maintenance and extension of Yamato II can be carried out on the library rather than the database quite efficiently in terms of labor and time; and (iii) browsing, editing and retrieval of the data can be performed without specifying the formats of the tables in the database. These three features make Yamato II not only flexible in changing and extending its functions but also operational with high fidelity.

DATA DISSEMINATION

The submitted data processed with Yamato II are now ready to be released to the public. The release process, however, is not necessarily automatic, because we have to observe the release dates the submitters have specified. There is no problem with releasing the data, if the submitters require an immediate release. When the paper reporting the sequence data is published, we override the specified date and immediately release the data. Sometimes, however, the submitter asks us to postpone the release of the data even in this situation, while some people urge us to release them. We can conjecture that the submitter and his/her colleagues are competing with other groups in the same area. At any rate, we have no responsibility for being involved in such a case, and go on to release the data without contacting the submitter. This is also common practice among the international databases.

Even if we operate this way, we have to keep a large amount of data unreleased until publication of the pertinent papers. These are called HUP (hold until published) data. Thanks to the international collaboration, information about publication of the papers is provided by GenBank, which has a tight connection to the MEDLINE service. If the submitter cues us to release the data while the paper is in press, however, the data will be opened to the public before the paper is published. Furthermore, if the submitted data are not intended to be published in a paper, the submitter's responsibility for cueing us will be heavier. It is noted that most EST data fall into this category. In this regard, it is crucial that the submitter specifies the date of release within a reasonable range at the time of submission. It is not us but the submitter who is primarily responsible for the data content and release.

DDBJ provides the following data retrieval services on WWW; keyword search, FASTA (4,5), BLAST (6,7), Malign (8), Clustal W (9). As shown in Figure 4, the WWW service at DDBJ has been extensively used for various purposes including data submission by Sakura, data retrieval in our databases, and linkage to other databases worldwide. FASTA and BLAST searches have also been frequently carried out. Though most of the accesses are made by Japanese users, another not minor portion is attributable to foreign accesses. Figure 5 represents the monthly increase in the number of users for DDBJ. Most of them are again Japanese researchers. The number has increased steadily and has exceeded 2000. We also provide other databases developed at NIG as shown in Table 1. If you are interested in these services, please refer to URL http://www.ddbj.nig.ac.jp/ .

RESEARCH ACTIVITIES

We believe that data banks themselves should be involved in doing research using the data they deal with, in order to provide high quality data. As mentioned above we also carry out a number of research projects in molecular evolution. In the following we will briefly introduce two of them which in particular require large scale data analysis.

Table 1. Other databases served at DDBJ

Database	Developer
Codon Usage Database (CUTG)[a]	T. Ikemura
Protein Mutant Database[b]	K. Nishikawa
C.elegans EST Database[c]	Y. Kohara
Bacillus subtilis Non-redundant Database[d]	G. Perriere and T. Gojobori

[a]Codon usage of individual genes among different species has been obtained from the international DNA sequence databases and presented (14).
[b]Protein mutations have been collected from relevant literature and presented with respect to positions and kinds of mutations and other features.
[c]EST sequences are retrievable in terms of gene names, chromosome numbers, clone names and others for *Caenorhabditis elegans.*
[d]All the duplications have been removed and all overlapping sequences have been merged for *Bacillus subtilis* genes and presented.
The database developers are all at the National Institute of Genetics except G. Perriere, who is at Lyon University in France.

First, we were interested in how prokaryotic genomes had been organized in the course of evolution. As a way to attack this problem, we compared the genomes of *Haemophilus influenzae* (10), *Mycoplasma genitalium* (11), *Escherichia coli* and *Bacillus subtilis* paying attention to the orthologous genes common to the four genomes (12). Positional comparison of orthologous genes we extracted revealed that the arrangement of the orthologous genes was quite species specific, indicating that the gene arrangement observed at present occurred after divergence among the four species. In particular, the gene arrangement in each species was so dynamic as to destroy operon structures even after the divergence between *H.influenzae* and *E.coli.* Though rarely, we could also find several conserved regions; the longest region includes the S10, spc and alpha operons as a linked unit common to the four species. We believe that strong selections have operated on those exceptional regions.

Second, if we extend the above discussion further, we would face the problem with not a genome but a gene; how a gene was evolutionarily constructed from its subregions. To study this problem, we extracted evolutionary motifs from as many protein sequences as possible (13). Actually, we chose the complete DNA sequences from the most recent international DNA sequence databases available then. We obtained 107 041 such sequences, and translated them into amino acid sequences. Those translated sequences were classified into subgroups in terms of sequence homology. The sequences of each subgroup were then aligned on the basis of their evolutionary relationships. For each group of aligned amino acid sequences, regions of evolutionarily conserved amino acid sites were searched for and extracted by window analysis. We called each of those regions the evolutionary motif, and they were 20–200 amino acids in length with 60 as the most frequent length. We note that many functional motifs and domains such as the homeo domain, POU specific domain, and bZIP motif are the size of ~60 amino acids. We speculate that DNA sequences corresponding to this size range have been used as building blocks for the present genes.

We believe that information biology will be one of the most important areas in biology, medicine and agriculture in the next century, and that molecular evolution will form a core in information biology. As mentioned earlier, a great number of genes and other DNA regions have been sequenced and have accumulated in the international DNA sequence databases. The biological functions of many sequences have, however, been unelucidated. The origins and functions of those genes and regions will be sought and solved by way of information biology in particular large scale data analysis using high performance computers. We are now in the position to analyze DNA and protein not only *in vivo* and *in vitro* but also *in silico.*

REFERENCES

1 Yamamoto, H. *et al*. In Akutsu, T. (ed.) *The Proceedings of Genome Informatics Workshop '96*, in press.
2 Adams, M.D. *et al.* (1991) *Science* **252**, 1651–1656.
3 Koike, T. *et al*. In Akutsu, T. (ed.) *The Proceedings of Genome Informatics Workshop '96*, in press.
4 Lipman, D.J. and Pearson, W.R. (1985) *Science* **227**, 1435–1441.
5 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
6 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol*. **215**, 403–410.
7 Gish, W. and States, D.J. (1993) *Nature Genet.* **3**, 266–272.
8 Hein, J. (1990) In Doolittle, R.F. (ed.) *Methods in Enzymology,* Academic Press, New York, Vol 183, pp. 626–645.
9 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) *Nucleic Acids Res.* **22**, 4673–4680.
10 Fleischmann, R.D. *et al.* (1995) *Science* **269**, 496–512.
11 Fraser, C.M. *et al.* (1995) *Science* **270**, 397–403.
12 Watanabe, H., Mori, H. and Gojobori, T. *J. Mol. Evol*. in press.
13 Tateno, Y. *et al. J. Mol. Evol.* in press.
14 Nakamura, Y., Wada, K., Wada, Y., Doi, H., Kanaya, S., Ikemura, T. and Gojobori, T. (1996) *Nucleic Acids Res*. **24**, 214–215.